APPENDIX A. EVALUATION OF EXISTING SQV SETS INCLUDING CHEMICAL DATA SETS

APPENDIX A. EVALUATION OF EXISTING SQV SETS INCLUDING CHEMICAL DATA SETS

A.1 RELIABILITY OF EXISTING SEDIMENT QUALITY VALUES

This appendix provides a detailed description of the methods and results of the reliability analysis for existing SQV sets in North America.

Five SQV sets already in use in North America were included in the reliability analysis (for a more complete description of the SQV sets evaluated, see Avocet and SAIC (2002) or the specific references cited below):

- **TELs/PELs –** TELs/PELs are derived using the database percentile method. TELs are intended to represent chemical concentrations below which biological effects rarely occur. PELs are intended to represent chemical concentrations above which adverse biological effects frequently occur. TELs/PELs were derived by classifying sediment samples within each data set as either toxic or non-toxic. TELs were calculated as the geometric mean of the 15th percentile of the effects distribution and the 50th percentile of the no-effects distribution. PELs were calculated as the geometric mean of the 50th percentile of the effects distribution and the 85th percentile of the no-effects distribution. TEL/PEL values have been developed for 8 metals, 12 individual PAHs, total PCBs, and 7 chlorinated pesticides (CCME 2002).
- **TECs/PECs –** Consensus-based SQVs have been proposed by a group of private and agency sediment researchers in an attempt to unify the wide variety of SQVs available in the literature (Ingersoll et al. 2000; MacDonald et al. 2000). Threshold effects concentrations (TECs) were derived using a group of existing freshwater SQV sets that represented levels below which adverse effects were seldom observed. TECs are considered conservative screening tools and not intended for use as cleanup goals. Similarly, probable effects concentrations (PECs) were derived using a group of existing freshwater SQV sets that represented levels above which adverse effects would be expected. If three or more published values with a similar narrative intent were available for a chemical or group of chemicals, the TEC or PEC was calculated as the geometric mean of these values. TECs and PECs have been developed for 8 metals, 10 individual PAHs, total PAHs, total PCBs, and 9 chlorinated pesticides (MacDonald et al. 2000).
- **LELs/SELs –** The screening level concentration approach was developed by the Ontario Ministry of the Environment and is based on the presence or absence of benthic species in freshwater sediments (Persaud et al. 1993). First, a field database of synoptic chemical and

benthic community data was compiled. A chemical concentration distribution was prepared for each benthic species and each chemical using only those stations at which each species was observed. For each distribution, the 90th percentile was determined. This concentration is assumed to represent a conservative estimate of the upper tolerance level for that species and that chemical since above that level the species is seldom observed. For each chemical, the tolerance levels of all the species are plotted on a graph by increasing concentration. From this distribution, various levels can be selected, depending on what percent of the species is to be protected. The most widely used values, developed by the Ontario Ministry of the Environment for use in the Great Lakes, include the "lowest effect level" (5th percentile) and the "severe effect level" (95th percentile). The LEL corresponds to a level at which you would expect to see effects in only 5% of benthic species, while the SEL represents a level at which you would expect to see effects in 95% of benthic species.

- **Washington Freshwater SQS/CSL –** The floating percentile method was developed in an effort to improve the reliability of freshwater SQVs for Washington State (Avocet 2003; Avocet and SAIC 2002). An optimal percentile of the data set that provides a low false negative rate is selected, and then each individual chemical concentration is adjusted upward until the false positive rate has decreased to its lowest possible level while retaining the same false negative rate. The method is designed to reduce mathematical error associated with the use of fixed percentiles for all chemicals. Sediment quality standards (SQS) and cleanup screening levels (CSLs) were calculated using the FPM for 11 metals, 16 individual PAHs, LPAHs, HPAHs, 4 phthalates, dibenzofuran, and total PCBs. These SQVs were derived using a large data set, primarily from western Washington and Oregon and including all of the Portland Harbor data that existed at that time (2001), and are currently applicable to freshwater sediments in Washington State (Avocet 2003).
- **Quotient Methods –** Quotient methods were developed as an approach to increase the predictive ability of certain SQVs described above (Long et al. 1998), and have been applied to TELs/PELs and TECs/PECs. Several quotient methods are available, some of which use individual metals and PAHs and others of which sum chemical classes. Based on the exploratory analysis conducted for this data set, several chemical classes such as PAHs and PCBs appeared to be more predictive of toxicity when summed. Therefore, quotients that use summed values, such as the mean PEL-Q, may be more appropriate. This is also the approach recently adopted for use in British Columbia (Macfarlane et al. 2002). However, it does not include all of the chemicals of interest at the site. Therefore, an

alternative version was also evaluated (SQG-Q) based on a recent paper by Fairey et al. (2001), which includes additional chemicals of interest, such as chlordanes and dieldrin.

For each existing SQV set, the more protective of the two thresholds (TEL, TEC, LEL, and SQS) was compared to the Level 1 and 2 biological effects levels, and the higher of the two thresholds (PEL, PEC, SEL, and CSL) was compared to the Level 3 biological effects levels, consistent with the narrative intent of these SQVs.

A.2 RELIABILITY ASSESSMENT METHODS

This section presents the methods used to obtain the appropriate chemistry data for the comparison of each SQV set and to evaluate the toxicity test endpoints. The chemistry data methods are presented in Section B.2.1; the toxicity data methods are presented in Section B.2.2.

A.2.1 Chemistry data methods

The project database was queried to obtain all chemistry data for the selected group of analytes (depending on the SQV set being evaluated), excluding any data qualified with a U, N, or R (see Section 2.2.1). To evaluate the reliability of existing SQV sets, chemical concentrations were summed in the same manner as that used in deriving each set of existing SQVs (e.g., threshold effects levels [TELs] and probable effects levels [PELs]) to facilitate comparison. For example, if the SQV set included values for individual PAHs, individual PAH concentrations were used in the reliability analysis. If the SQV set used low-molecular-weight PAH (LPAH) and high-molecular-weight (HPAH) sums, these sums were used instead.

These data were downloaded into Microsoft Excel® files, which are included in this appendix. There are 15 Excel® files, one for each combination of the three effects levels and five endpoints (four individual endpoints and one pooled endpoint). For SQV sets other than the PEL-Qs, the following approach was used. The first worksheet, entitled "BioHits," contains the biological hit/no-hit results for the endpoint and effects level being evaluated. The worksheet "ChemData" shows the chemistry data for all stations downloaded from the SEDQUAL Information System, organized by chemical and increasing concentration. A Visual Basic® macro called MakeTable is then run to organize the data into a data table, shown in the worksheet DataTable. The DataTable worksheet also has a column into which the biological hit/no-hit values are entered for each station. Blank cells indicate analytes for which no data are available at those stations. The reliability macro skips these cells.

The final worksheet, entitled Criteria, contains the individual SQVs for each of the four SQV sets that are being assessed for the 34 analytes included among the various SQV

sets.[1] These values are pre-entered in columns H-AO of the worksheet. To the left of these values, there are columns for each of the seven measures of reliability, which are calculated by a Visual Basic® macro called TestReliability. The TestReliability macro compares the chemical concentrations of each chemical at a station to the corresponding SQVs and determines whether a hit or no-hit would be predicted at that station. Then the chemical hit/no-hit prediction is compared to the biological hit/no-hit value, and the macro records whether the result is a correct prediction, a false positive, or a false negative. From these results, each of the other reliability parameters was calculated. These and the other Excel® macros were manually verified to ensure their accuracy. The seven reliability parameters are listed below:

- **False negatives –** Incorrectly predicted no-hits/total hits
- **False positives –** Incorrectly predicted hits/total no-hits
- **Sensitivity –** Correctly predicted hits/total hits
- **Efficiency –** Correctly predicted no-hits/total no-hits
- **Predicted hit reliability –** Correctly predicted hits/total predicted hits (this measure is equivalent to "1988 Efficiency" in Avocet (Avocet 2003; Avocet and SAIC 2002))
- **Predicted no-hit reliability –** Correctly predicted no-hits/total predicted no-hits
- **Overall reliability –** Correctly predicted stations/total stations

For the quotient methods, the chemistry was downloaded, and both the probable effects level quotient (PEL-Q) and the sediment quality guideline quotient (SQG-Q) were calculated for each station. The PEL-Q was calculated for each sediment sample by summing the average quotient for seven metals (i.e., arsenic, cadmium, chromium, copper, lead, mercury, and zinc), the quotient for total PAHs, and the quotient for total PCBs and then dividing this sum by three. The SQG-Q used the sum of the quotients of each individual chemical or class included in the equation, divided by the number of chemicals or classes, and was calibrated using an empirical approach in which a variety of different equations was tested using various possible SQGs as the basis for the quotient. The chemicals included, and the SQGs on which their quotients are based, are: cadmium (PEL), copper (effects range median [ERM]), silver (PEL), lead (PEL), zinc (ERM), total chlordane (ERM), dieldrin (ERM), total PAHs (PEC), and total PCBs (PEC). PAHs are also OC-normalized in this approach.

A.2.2 Toxicity data methods

Two endpoints, growth and mortality, were included in the reliability assessment. The mortality endpoint was obtained for both toxicity tests at all 233 stations, whereas the growth endpoint could not be obtained for a few stations because of 100% mortality in

[1] The macros for the spreadsheets were set up using the word "criteria." However, for the Portland Harbor project, the word "criteria" should be replaced with the word "SQV."

the same samples. The types and numbers of toxicity test endpoints in the Round 2 data set are summarized in Table A-1.

Table A-1. Round 2 toxicity tests and endpoints

Test	Maximum Number of Stations[a]
Hyalella azteca	
28-day mortality	233
28-day growth	229
Chironomus tentans	
10-day mortality	233
10-day growth	227

[a] Some of the stations may have been labeled "Indeterminate" for one or more of the effects levels. The number of endpoints directly correlates to the number of stations.

For the reliability assessment, each of the four individual endpoints was assigned to the three biological effects levels based on the definitions stated in Section 2.2.3. In addition, a pooled endpoint was derived by combining all four endpoints from the two tests. Table A-2 shows the number and percentage of stations associated with biological hits for each effects level and endpoint combination.

Table A-2. Biological hits

Effects Level	Number of Biological Hits (percent)[a]				
	***Chironomus* growth**	***Chironomus* mortality**	***Hyalella* growth**	***Hyalella* mortality**	**Pooled endpoint[b]**
Level 1	29 (13%) [12]	47 (21%) [11]	139 (66%) [18]	30 (13%) [3]	167 (78%) [18]
Level 2	24 (11%) [0]	34 (15%) [0]	98 (43%) [0]	20 (9%) [0]	128 (55%) [0]
Level 3	17 (7%) [0]	25 (11%) [0]	46 (20%) [0]	18 (8%) [0]	77 (33%) [0]

[a] The denominator used to determine the percentage of hits excludes the number of statistically indeterminate samples shown in brackets.

[b] For this analysis, all four biological endpoints were combined into a single pooled endpoint. For later analyses, biological endpoints were pooled by species.

As can be noted from Table A-2, there were substantial differences among endpoints in the observed responses. The *Hyalella* growth test showed a response at a greater number of stations than any of the other toxicity test endpoints for all effects levels. The *Chironomus* growth test was comparable to the *Hyalella* mortality test in the number of adverse responses exhibited at each effects level; they both exhibited the fewest number of responses among the endpoints. *Chironomus* mortality was intermediate in the number of responses exhibited at each effects level. The pooled endpoint always

exhibited a response at a relatively large number of stations as compared to any one individual endpoint, suggesting that there were frequent differences in the endpoints exhibiting effects among stations.

A.3 RELIABILITY ANALYSIS

The reliability analysis for each of the effects levels is discussed in this section. To simplify the discussion, the evaluation below focuses on the four primary reliability parameters: sensitivity, efficiency, predicted no-hit reliability, and predicted hit reliability. Two of the other parameters, false positives and false negatives, are simply 100% minus sensitivity and efficiency. The final parameter, overall reliability, is less useful in this analysis because it is dependent on the proportion of hits to no-hits in the data set, which varies significantly among effects levels.

A.3.1 Level 1

Table A-3 presents the results for the four SQV sets that were assessed at Level 1. The TEL, TEC, and LEL levels all performed similarly and very conservatively, although in general, the TECs performed 10 to 15% better with respect to efficiency than the TELs and LELs. In all three cases, the SQV sets had very high sensitivity (few false negatives). On the other hand, these SQV sets classified nearly every sample as a hit, leading to a very high false positive rate (100% in the case of the TELs). In general, these SQV sets predicted that all or nearly all samples would be hits, and the proportion of correctly predicted hits simply reflects the proportion of actual biological hits in the data set. Therefore, these SQV sets are not really useful in making correct predictions about lower effects levels. Although it is highly likely that any sample with chemical concentrations that fall below these levels will not exhibit biological effects, there will be few to no samples with chemical concentrations that are that low. Relatively large, apparent variations in the predicted no-hit reliability parameter actually represent only a few samples, inasmuch as very few samples overall are predicted to be no-hits.

Table A-3. Reliability analysis for Level 1 biological effects

SQV Set	% Sensitivity	% Efficiency	% Predicted Hit	% Predicted No-Hit
***Chironomus* Growth**				
TEL	100	10	13	100
TEC	100	23	14	100
LEL	97	10	12	67
Washington SQS	83	51	17	91
***Chironomus* Mortality**				
TEL	98	7	20	67
TEC	94	20	22	90
LEL	96	6	20	33
Washington SQS	68	47	23	81
***Hyalella* Growth**				
TEL	98	23	59	na
TEC	88	34	60	34
LEL	99	26	60	67
Washington SQS	60	54	60	31
***Hyalella* Mortality**				
TEL	98	2	13	67
TEC	85	15	14	93
LEL	98	2	13	33
Washington SQS	57	43	15	89
Pooled Endpoint				
TEL	98	27	71	na
TEC	90	42	73	34
LEL	99	29	72	33
Washington SQS	63	61	75	23

na – did not predict any no-hits at this effects level

The Washington State freshwater SQS values are less conservative than the other three SQV sets. While they have 20 to 40% higher efficiency, it comes at the expense of 20 to 40% lower sensitivity, particularly for the more sensitive 28-day *Hyalella* endpoints, which were not included in the original calculation of these SQVs due to the lack of sufficient data at that time. These SQVs likely need to be recalculated to take into account the chronic bioassay data in order to obtain better performance with this data set.

A.3.2 Level 2

Table A-4 shows the reliability results for Level 2, which are overall very similar to those of Level 1. Again, the TEL, TEC, and LEL SQVs all classify nearly all samples as hits, resulting in high sensitivity and very low efficiency. The predicted hit and predicted no-hit reliability values appear different from those of Level 1; but in reality, these values just reflect the fact that there are fewer actual hits at Level 2, especially for

the *Hyalella* toxicity test endpoints. Therefore, the predicted hit reliability declines because most samples are still predicted to be hits. For the Washington freshwater SQS values, the same pattern is observed – sensitivity and efficiency are nearly the same as those at Level 1, while predicted hit reliability declines because there are fewer biological hits at this level, especially in the *Hyalella* test.

Table A-4. Reliability analysis for Level 2 biological effects

SQV Set	% Sensitivity	% Efficiency	% Predicted Hit	% Predicted No-Hit
***Chironomus* Growth**				
TEL	100	4	10	100
TEC	100	17	12	100
LEL	96	4	10	67
Washington SQS	83	46	14	96
***Chironomus* Mortality**				
TEL	100	2	15	100
TEC	97	14	16	97
LEL	97	1	14	67
Washington SQS	76	43	19	91
***Hyalella* Growth**				
TEL	99	4	42	67
TEC	92	19	44	72
LEL	100	5	42	100
Washington SQS	62	45	43	61
***Hyalella* Mortality**				
TEL	100	1	9	100
TEC	100	14	10	100
LEL	95	1	8	67
Washington SQS	80	42	12	96
Pooled Endpoint				
TEL	99	2	55	67
TEC	94	20	59	72
LEL	99	2	55	67
Washington SQS	66	49	61	54

A.3.3 Level 3

The reliability results for Level 3 are presented in Table A-5. Most of the SQV sets appear to perform better at this effects level, with a few exceptions (notably a lack of sensitivity in comparison to the *Hyalella* growth results). At this level, the Washington CSLs come more into line with the other SQV sets, tending to be most similar to the PELs in performance. Among all the SQV sets, there is a better balance between sensitivity and efficiency, although judging by the low predicted hit reliability values, there is still a tendency to over-predict actual hits by a substantial amount (three times the actual number of hits).

Table A-5. Reliability analysis for Level 3 biological effects

SQV Set	% Sensitivity	% Efficiency	% Predicted Hit	% Predicted No-Hit
***Chironomus* Growth**				
PEL	82	59	13	97
PEC	65	70	14	95
SEL	53	80	16	95
Washington CSL	65	54	9	95
***Chironomus* Mortality**				
PEL	68	57	16	94
PEC	56	68	17	93
SEL	52	79	23	93
Washington CSL	72	53	16	94
***Hyalella* Growth**				
PEL	44	56	19	80
PEC	31	66	17	79
SEL	31	80	25	82
Washington CSL	51	52	20	81
***Hyalella* Mortality**				
PEL	72	56	12	96
PEC	67	68	15	96
SEL	67	79	21	97
Washington CSL	83	53	13	97
Pooled Endpoint				
PEL	57	59	40	74
PEC	45	70	42	72
SEL	41	84	55	74
Washington CSL	61	55	40	74

A.3.4 Quotient method

Pooled results for the SQG-Q and PEL-Q methods are shown in Figures A-1 and A-2, respectively. The x-axes present the full range of quotient values (SQG-Q and PEL-Q), and the y-axes present the percentage of hit classification. At each level of effects, a full range of possible quotients was evaluated to determine if there was a quotient level that could reliably predict hits and no-hits in the data set. The pink line shows the percentage of no-hits below the quotient value, while the blue line shows the percentage of hits above the quotient value. Ideally, both levels would be high (e.g., above 80%) in order for a selected quotient value to have good reliability in predicting both hits and no-hits. As can be seen from the graphs, this does not occur at any effects levels throughout the range of possible quotient values, except in some cases at the extreme ends of the data distribution. Setting values at the ends of the distributions would not be helpful because only a few stations fall below these levels (at the low end) or above these levels (at the high end).

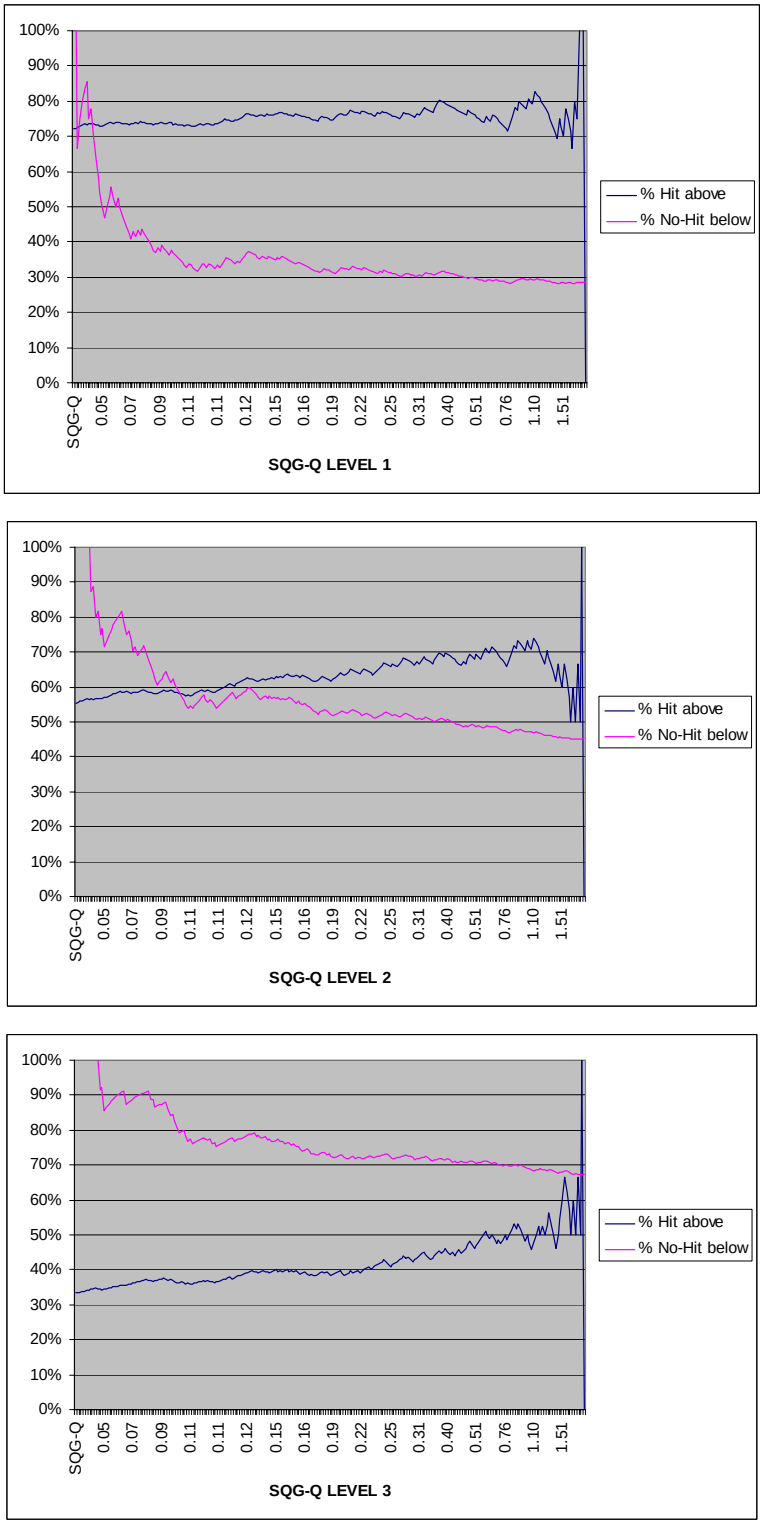

Figure A-1. SQG-Q pooled endpoint hit and no-hit screening curves

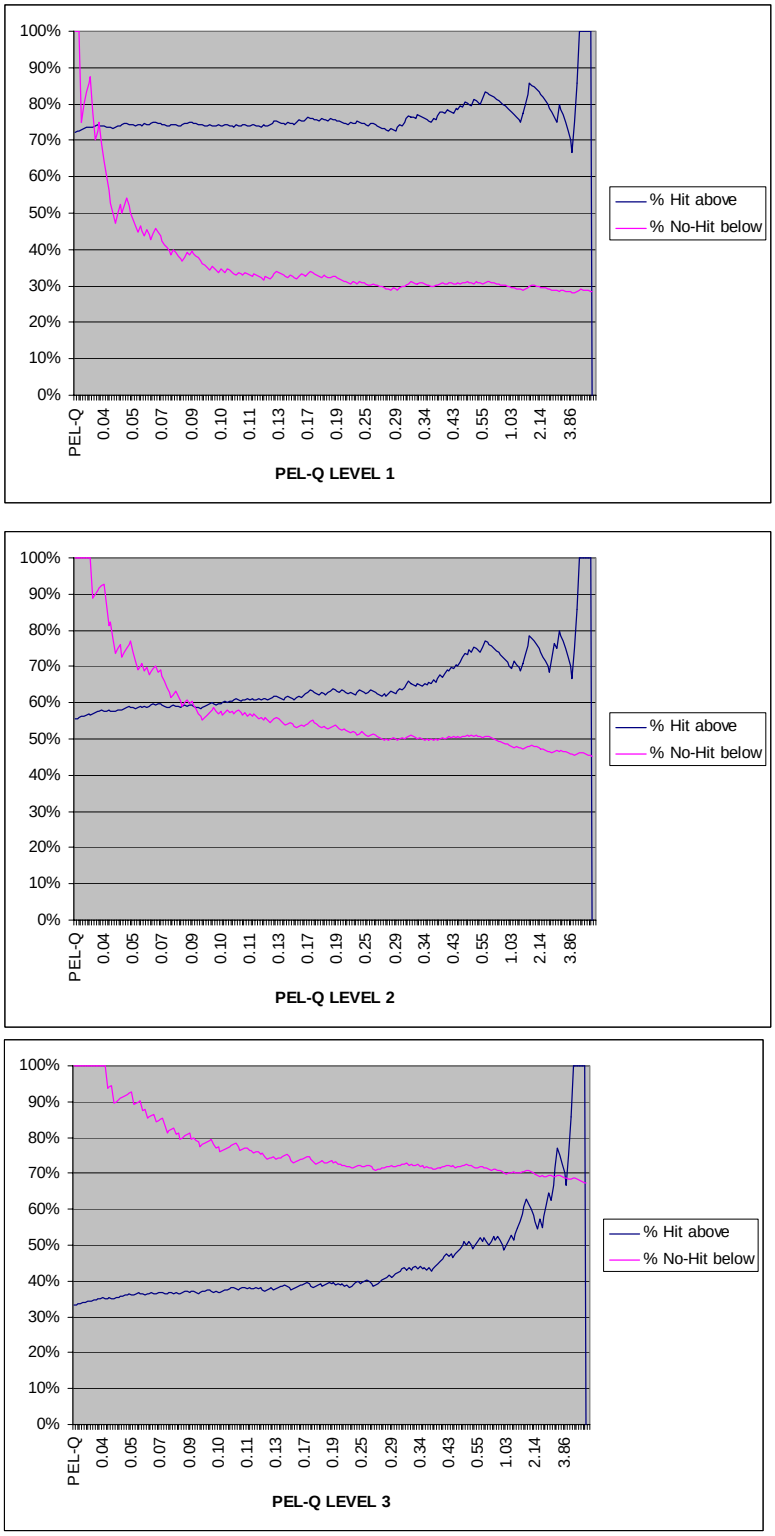

Figure A-2. PEL-Q pooled endpoint hit and no-hit screening curves

Even though a single quotient value may not be reliable for predicting both hits and no-hits, lower levels could be used to screen out areas (identify no-hits), and higher levels could be used to screen in areas (identify hits). Unfortunately, this approach also has very low reliability. At Level 1, the no-hit screening (the pink line) has a reliability of only about 30 to 40% across most of the distribution. At Level 3, the hit screening (the blue line) has only about 40% reliability through most of the data set, rising to 60% near the upper end. The intermediate Level 2 effects level has the best balance of reliability for both quotient measures but only achieves about 60% reliability for both hit and no-hit screening.

In general, this is an improvement over most of the SQV sets discussed above although not sufficiently reliable for use in predicting toxicity results at this site. It is possible that the quotient approach has merit, but it needs to be optimized on a site-specific basis. Both of the quotient methods tested here were developed based on data sets for marine and estuarine waters throughout the United States. The PEL-Q quotient method was specifically optimized for predicting acute amphipod toxicity in the data set used to develop the PEL-Q and therefore may not be optimal for the Portland Harbor data set, because it is clear that different chemicals are affecting different endpoints.

A.4 SUMMARY OF RELIABILITY RESULTS FOR EXISTING SQV SETS

None of the existing SQV sets perform well enough to use them in predicting biological effects at the Portland Harbor Superfund Site. The lower thresholds (the TELs, TECs, and LELs) are far too conservative to be useful because they classify all or nearly all stations as hits (low efficiency). The higher thresholds (the PECs, PELs, and SELs) are more successful at predicting toxic effects. None of the existing SQV sets perform well enough to use them in predicting biological effects at the Portland Harbor Superfund Site. The lower thresholds (the TELs, TECs, and LELs) are far too conservative to be useful because they classify all or nearly all stations as hits (low efficiency). The higher thresholds (the PECs, PELs, and SELs) are more successful at predicting toxic effects, yet the error rates are still high enough that substantial portions of the Study Area could be incorrectly classified as contributing to adverse effects.

Error rates are still high enough that substantial portions of the Study Area could be incorrectly classified as contributing to adverse effects. It is possible that the development of a site-specific SQV set or predictive model could reduce error rates.

A.5 REFERENCES

Avocet. 2003. Development of freshwater sediment quality values for use in Washington State. Phase II report: Development and recommendation of SQVs for freshwater sediments in Washington State. Publication No. 03-09-088. Prepared for Washington Department of Ecology. Avocet Consulting, Kenmore, WA.

Avocet, SAIC. 2002. Development of freshwater sediment quality values for use in Washington State. Final report. Publication No. 02-09-050. Prepared for Washington Department of Ecology. Avocet Consulting, Kenmore, WA and Science Applications International Corporation (SAIC), Bothell, WA.

CCME. 2002. Canadian sediment quality guidelines for the protection of aquatic life: summary tables. Canadian Environmental Quality Guidelines, 1999, updated 2001, updated 2002 [online]. Canadian Council of Ministers of the Environment. Updated 2002. [Cited March 2005]. Available from: http://www.ccme.ca/assets/pdf/e7_002.pdf.

Fairey R, Long ER, Roberts CA, Anderson BS, Phillips BM, Hunt JW, Puckett HM, Wilson CJ. 2001. An evaluation of methods for calculating mean sediment quality guideline quotients as indicators of contamination and acute toxicity to amphipods by chemical mixtures. Environ Toxicol Chem 20(10):276-2286.

Ingersoll CG, MacDonald DD, Wang N, Crane JL, Field LJ, Haverland PS, Kemble NE, Lindskoog RA, Severn C, Smorong DE. 2000. Prediction of sediment toxicity using consensus-based freshwater sediment quality guidelines. US Environmental Protection Agency, Chicago, IL.

Long ER, Field LJ, MacDonald DD. 1998. Predicting toxicity in marine sediments with numerical sediment quality guidelines. Environ Toxicol Chem 17(4):714-727.

MacDonald DD, Ingersoll CG, Berger TA. 2000. Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. Arch Environ Contam Toxicol 39:20-31.

Macfarlane MW, MacDonald DD, Ingersoll CG. 2002. Criteria for contaminated sites: criteria for managing contaminated sediment in British Columbia. Technical appendix. Ministry of Water, Land and Air Protection, Victoria, BC, Canada.

Persaud D, Jaagumagi R, Hayton A. 1993. Guidelines for the protection and management of aquatic sediment quality in Ontario. ISBN 0-7729-9248-7. Ontario Ministry of Environment and Energy, Toronto, ON.